NIST'S ASSESSMENT OF TEXT INDEPENDENT SPEAKER RECOGNITION PERFORMANCE

Mark Przybocki, Alvin Martin

National Institute of Standards and Technology
mark.przybocki@nist.gov, alvin.martin@nist.gov

ABSTRACT

NIST has coordinated annual evaluations of text-independent speaker recognition since 1996. These evaluations aim to provide important contributions to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text-independent speaker recognition.

The evaluations have focused primarily on speaker detection in the context of conversational telephone speech. The evaluations are designed to foster research progress with the objectives of exploring promising new ideas in speaker recognition, developing advanced technology incorporating these ideas, and measuring the performance of this technology.

Evaluation participants have included commercial, academic and governmental research laboratories from around the world. This paper reviews how NIST assesses speaker recognition systems through our series of benchmark evaluations, focusing on the 2002 NIST Speaker Recognition evaluation.

1. INTRODUCTION

The National Institute of Standards and Technology (NIST) Speech Group has been coordinating evaluations of language technologies since 1987. From the early days of speaker dependent Resource Management automatic speech recognition tests [1], where the task was to create automatic transcripts from audio (speech-to-text) of read sentences modelled after a naval resource management task, such as: "*Show me the maximum speed for vessels in the Bering Strait*", to the current more complex tasks of speaker independent speech-to-text combined with meta-data extraction from a continuous audio stream [2], assessment was recognized as a primary activity for driving the technology forward.

By providing explicit evaluation plans (which specify the evaluation protocols), common tests sets, standard measurements of error, tools for data manipulation and a forum for openly discussing algorithm successes and failures, NIST has led the way in providing a series of benchmark tests for automatic speech recognition, language identification, topic detection and tracking, automatic content extraction, spoken document retrieval, machine translation and speaker recognition.

NIST serves the role of coordinating the speaker recognition (and other) evaluations. NIST designs the tests for local implementation by all participants. Only their results are returned to NIST for scoring. The reported results are not to be construed, or represented, as endorsement of any participant's system, or as official findings on the part of NIST or the U.S. Government.

2. SPEAKER RECOGNITION TASKS

Four types of tasks have been included in some of the annual NIST Speaker Recognition evaluations.

- One-Speaker Detection: Task is to determine whether a specified speaker is speaking in a given single-channel segment of speech.
- Two-Speaker Detection: Task is to determine whether a specified speaker is speaking in a given summed two-channel segment of speech.
- Speaker Tracking: Task is to perform speaker detection as a function of time. Systems are required to identify the time intervals (if any) in which a known speaker is speaking in a summed two-channel segment of speech.
- Speaker Segmentation: Task requires a system to perform speaker clustering. All segments of speech must be associated with one or more unknown speakers.

Of these four tasks, it is the one-speaker detection task that has been a part of each evaluation and is the one most central to biometric identification using speech. Recent evaluations have introduced a variant of the one-speaker detection task, referred to as the "extended data" test. It should be noted that although it is a different test, the assessment procedures remain the same.

While the other three tasks (two-speaker detection, speaker tracking, and speaker segmentation) have had an important place in NIST Speaker Recognition evaluations, they are not discussed here.

2.1 One-Speaker Detection

This is the basic speaker recognition task that has been part of all the NIST Speaker Recognition evaluations. The task is to determine whether a specified speaker is speaking in a given single-channel segment of mu-law encoded telephone

speech. The hypothesized speakers are always of the same sex as the segment speaker (the speaker in the test segment).

The task each year consists of a sequence of trials; the main one-speaker test in 2002 had about 39,000 trials. A trial consists of a single hypothesized speaker and a specific test segment. The system is required to make an actual decision (true or false) on whether the specified speaker is present in the test segment. Along with each actual decision systems are required to provide for each trial a likelihood score indicating the degree of confidence in the decision. A trial where the hypothesized speaker is present in the test segment (correct answer true) is referred to as a target trial. Other trials (correct answer false) are referred to as impostor trials or non-target trials.

The actual decisions and likelihood scores are used as the basis for evaluating system performance.

2.2 The Extended Data Task

In 2000, George Doddington [3] suggested a radically different approach to the one speaker detection, which later became known as the "extended data" task. Doddington observed that people do a better job of detecting those with whom they are quite familiar than those they do not know well. They become accustomed to the speaking habits and idiosyncrasies of those they know well. He suggested making use of idiolectal characteristics of speakers for whom considerable transcribed speech data was available. Doddington showed that by using the available manual transcripts of the Switchboard 1 corpus, one could make use of the word patterns – specifically the common unigrams, bigrams, and trigrams – of individual speakers for detection purposes.

In 2001 the first extended data task was offered as a dry-run evaluation. The underlying task is the same as the one-speaker detection task with the only differences being the amount of data used to train a speaker model, the duration of the test segments, and a provision encouraging the use of automatic speech recognition transcripts, supplied for all of the training and test data.

After a successful dry-run evaluation using the original release of Switchboard, a formal extended data evaluation was offered in 2002, using two phases of Switchboard II (phases 2 and 3).

3. DATA

The primary data sources for the NIST Speaker Recognition evaluations have been the Switchboard Corpora of conversational telephone speech collected over the last decade by the Linguistic Data Consortium (LDC). These all involve five to ten minute conversations between two speakers. The speakers are paired and assigned a conversational topic by an automatic system. They are recruited adults generally paid nominal fees for their participation. Speaking on-topic has sometimes been optional. The collections of Switchboard (SWBD) are documented in Table 1.

3.1 Training Data

Each evaluation test kit includes training data for every hypothesized speaker (referred to as model speakers). Through the course of the evaluations the amount of training data has generally remained at two minutes per model speaker. Early evaluations were designed to examine how varying the training data affects performance. These tests revealed that a large performance gain was achieved as the training data became more varied, either by including data from more than one conversation, or more than one telephone handset. These tests also showed that, not surprisingly, more training data (more than 2 minutes) also improves performance, but to a lesser degree.

Catalog #	Title	Conversations /Speakers	Attributes
LDC97S62	SWBD I	4870 / 543	U.S.A.
LDC98S75	SWBD II phase 1	3702 / 661	Mid-Atlantic
LDC99S79	SWBD II phase 2	4575 / 684	Mid-West
LDC2002S06	SWBD II phase 3	2728 / 640	South
LDC2001S13	SWBD cellular 1	1309 / 254	East Coast GSM dominate
not available	SWBD cellular 2	2020 / 419	East Coast varied transmissions

Table 1: The Switchboard Corpora, all two-channel mu-law encoded data of conversational telephone speech in American English, available from the LDC [4].

3.1.1 One-speaker detection training

Since 2000, training data for the NIST Speaker Recognition evaluations has been "one-session" training, where two minutes of training data is supplied from one single conversation. This choice has the benefit of using fewer conversations for training speaker models, and therefore more conversations are available for testing. See section 5.5.1, Training Conditions, for performance by varying training conditions.

In all of the NIST Speaker Recognition evaluations, training data has been created by concatenating consecutive turns of speech of the model speaker. Areas of silence are removed. The training data is taken from the tail end of a conversion, which may contain more natural conversational speech, avoiding the sometimes-awkward introductions that occur when two strangers are conversing for the first time.

3.1.2 Extended Data training

The extended data task has sought to use higher levels of information for speaker recognition. Whole conversation sides were used to train each hypothesized speaker. There were five training conditions, differing only by the number of conversation sides used to train each speaker model (1, 2, 4, 8, or 16).

Systems were allowed to use both the acoustic data and the NIST supplied ASR transcripts [5] to train each speaker model.

3.2 Test data

The first few NIST Speaker Recognition evaluations contained tests segments of fixed durations (3, 10 and 30 seconds lengths). These evaluations confirmed the intuitive expectation that longer test segments improve performance, and also established benchmarks for the amount of improvement.

3.2.1 One-speaker detection test data

Since 1999, NIST has supplied random length test segments that varied from just a few seconds up to one minute, but averaged about thirty seconds over the entire test set. They have been selected by choosing a minute from a conversation and concatenating the turns of each speaker within that minute into two separate test segments, one per channel. As with the training segments, areas of silence are removed and whole turns are included to the extent possible. No more than one test segment is created from each conversation side, and no test segments come from conversations that were used for training data.

3.2.2 Extended Data test data

A test segment for the extended data task has consisted of one whole conversation side.

In order to maximize the use of the limited data, a jackknifing procedure was used to make use of all the conversation sides as test segments with multiple models trained for each speaker using 1, 2, 4, 8 or 16 sides as a speaker's training data. For each training condition (number of sides), all of a speaker's conversation sides were used as a test segment exactly once.

3.3 Test Trials

A system is tested on each trial, and each test is made independently of all others. That is, the system under test must make its decision with knowledge only of the hypothesized model speaker and the test segment. Normalization over multiple test segments, or multiple model speakers, is not allowed.

3.3.1 One-speaker test trials

In general, each test segment is used in eleven separate trials, one of which is a target trial with the segment speaker being the model speaker. The other ten model speakers are selected from among all model speakers of the same sex as the segment speaker. This 10 to 1 ratio of impostors to target trials is NOT intended to reflect what is likely in an actual application environment. It does, however, serve to approximately minimize the variance of the primary metric discussed in section 4.

3.3.2 Extended data test trials

A given number of conversation sides are used as training for a model speaker (either 1, 2, 4, 8, or 16). Each remaining conversation side for the model speaker is tested only once for a given training condition. There are four non-target trials per speaker model, two of which are cross gender trials.

4. ASSESSMENT REPRESENTATION

The two types of errors that can occur in a detection task are denoted as missed detections and false alarms. The miss rate ($P_{Miss|Targ}$) is the percentage of target trials decided incorrectly. The false alarm rate ($P_{FA|NonTarg}$) is the percentage of impostor trials decided incorrectly. These error probabilities are determined from a system's actual decisions.

NIST has chosen to use a cost function defined as a weighted sum of the two types of errors as the basic performance measure. This cost, referred to as the C_{Det} cost, is defined as:

$$C_{Det} = (C_{Miss} * P_{Miss|Targ} * P_{Targ}) + (C_{FA} * P_{FA|NonTarg} * P_{NonTarg})$$

The required parameters in this function are the cost of a miss (C_{Miss}), the cost of a false alarm (C_{FA}) and the a priori probability of a target speaker (P_{Targ}). $P_{NonTarg}$ is then defined to be $1-P_{Targ}$.

For assessing speaker recognition systems, NIST has been using the following parameters: $C_{Miss}=10$, $C_{FA}=1$, $P_{Targ}=0.01$.

Unlike the 10 to 1 ratio used in defining the target to impostor trials that doesn't reflect any application use, a 10 to 1 penalty rate for misses over false alarms may be realistic for many applications. One advantage of this type of error metric formulation is that the test data need not resemble the intended application data in terms of target richness.

It is realistic to expect that a system without any knowledge of the speakers should have an expected cost of one. Such a system would either always decide that a trial was the target or conversely it would always decide that the trial was not the target. For the parameters that NIST has used, such a system that decided false for every trial, incurring a miss for all target trials, would be given a C_{Det} value of 0.1. While a system that decided true for every trial, incurring a false alarm for all non-target trials, would be given a C_{Det} value of 0.99. Therefore NIST has normalized the C_{Det} cost by the factor 0.1. Thus, a single number represents the C_{Det} cost.

More informative is a representation that shows all operating points for the system rather than just one. All operating points can be determined because each system is required to output a likelihood score with each decision of true and false. NIST produced scoring software [6] sweeps through these likelihood scores varying the threshold for whether the system would have chosen an actual decision of true or false, producing all possible operating points.

NIST has been using a variant of the popular receiver operating characteristic (ROC) curve as suggested by Swets [7], where the two types of error are plotted on the x and y axis using a normal deviate scale. NIST has termed this representation a detection error tradeoff (DET) curve [8]. DET curves have the key property that if the underlying distributions of scores for both targets and non-target trials are gaussian, then the resulting performance curve is a straight line. DET plots make it easier to view the separation between systems that are approaching very good performance. In the NIST evaluations the performance curves have almost always been close to linear.

Since the actual decision operating point is a particular point on the DET curve, this point can be plotted with a special symbol. One other point that NIST often plots on the DET curve is the point where the system has the lowest C_{Det} value. NIST refers to this operating point as the minimum C_{Det} point.

The second view of error that NIST uses is in the form of stacked bar charts. The stacked bar charts are used to show what portion of the C_{Det} value is due to miss detections and what portion of the value is due to false alarms. This is shown for both the actual decision and minimum C_{Det} points.

Examples of each, a DET curve and stacked bar chart, are shown in figures 1 and 2.

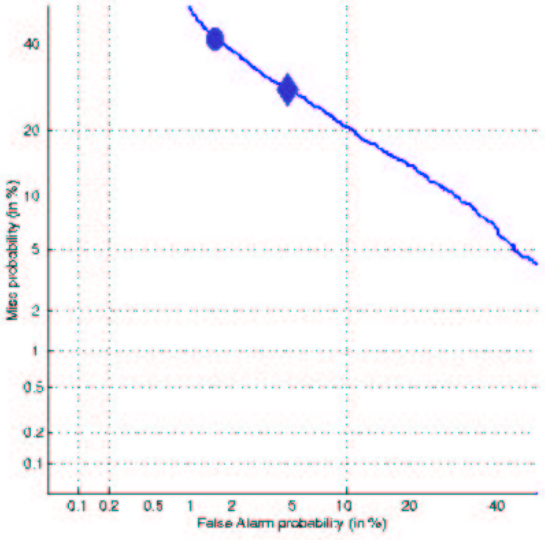

Figure 1: Example DET Curve showing speaker recognition performance. The Actual Decision C_{Det} value is marked with a diamond, and the Minimum Cost C_{Det} value is marked with a circle.

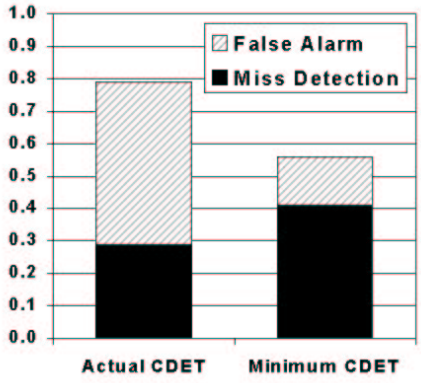

Figure 2: Example Stacked Bar Chart showing the error distributions of missed detections and false alarms of the Actual Decision C_{Det} and Minimum Cost C_{Det}.

5. ASSESSING EVALUATION RESULTS

A given test for a NIST speaker recognition task usually contains hundreds of speakers, with each speaker being used for several target and non-target trials. Although NIST reports results for the entire test set, which may demonstrate robustness of the systems over a variety of conditions, we also perform conditional analyses over subsets of the evaluation data to more closely analyze factors that affect performance. Some of these conditional analyses are described below, where we describe the conditional analysis using results from the most recent evaluations.

5.1 Primary Condition

The evaluation plan explicitly states the primary condition of interest. This condition consists of a subset of all trials. Over the evaluation series, the primary conditions have focused on a variety of conditions. It is important to define the primary condition before the evaluation test data is created, because this gives NIST a chance to tailor the test kit to include the primary condition in large numbers of trials and speakers. For instance, in 1998 the primary condition was defined to be "same number tests from electret handsets". Since participants received phone calls at a single number (usually at home or at work) and they were required to initiate each of their calls from a different phone, it was important to choose the training data from a conversation that they received. And in 1999 when the primary condition was "different number tests from electret handsets", it was important to choose an initiated call (from an electret handset) for training, making all tests against their other conversation sides, different number tests. Both of these scenarios maximized the number of trials available for the condition of interest.

The primary condition of the 2002 evaluation was defined as those trials where both the model speaker and the segment speaker were recorded over a cellular transmission and the amount of speech by the segment speaker was in the range of 15 to 45 seconds.

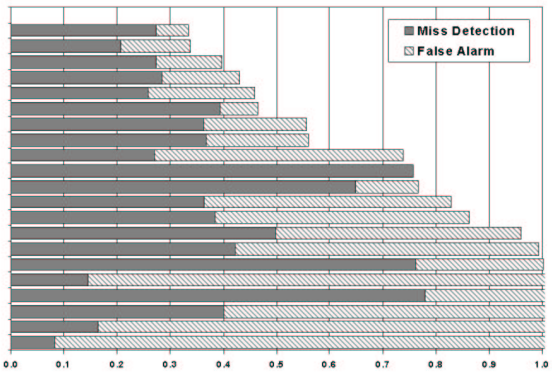

Figure 3: The stacked bar chart showing the actual decision C_{Det} values for 21 participating systems for the 2002 primary condition in the one-speaker detection task.

Figure 3 shows the stacked bar chart for the systems that participated in the 2002 one-speaker detection evaluation, when limiting the results to the primary condition. Figure 4 shows the corresponding DET plot.

Cleary one can see that there is a great deal of variation in the system performance for this task. Minimum C_{Det} costs range from .33 to over 1.0. In accord with NIST's understanding with the participants, we do not identify the various systems under test to the general public. Although the evaluations are open to all who want to participate, once the evaluation takes place the comparative results are only discussed at the follow-up meeting, after which sites are free to do what they wish with their own results, but can neither redistribute nor publish results from any other site without that site's express permission.

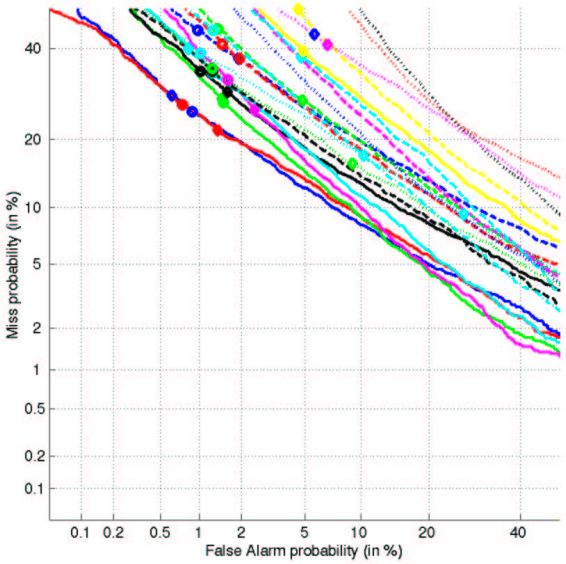

Figure 4: DET plot showing the results for 21 participating systems for the 2002 primary condition in the one-speaker detection task.

5.2 Duration of Test Segments

The NIST evaluations from 1996-1998 offered tasks with three categories of test duration, nominally 3, 10 and 30 seconds, where each 3 second test segment was a subset of speech in a 10 second segment, and each 10 second segment was a subset of a 30 second segment. Not surprisingly, these evaluations revealed the level of improvement that was obtained with increasingly longer test segments.

Recent evaluations contained varying length test segments, from nominally one second duration to almost sixty seconds, averaging around thirty seconds for the entire test set. The conditional analysis for test segment duration was performed by categorizing each test segment, based on the amount of speech by the segment speaker.

The five categories for duration were: 0-15+ seconds, 16-25+ seconds, 26-35+ seconds, 36-45+ seconds, and greater than 46 seconds. Figure 5 shows the DET curves for a typical well-performing system from the 2002 evaluation for the duration condition. These results indicate that performance is significantly lower for segments that are shorter than 15 seconds, but that performance is not greatly affected for segments that are longer than 15 seconds. Although consistent with the findings in years when test segments were explicitly falling into three categories, current duration analysis also indicates that the duration effect seen is limited and that once some minimum duration (apparently in the 10 to 15 second range) is available, the amount of test speech ceases to be a major factor in performance.

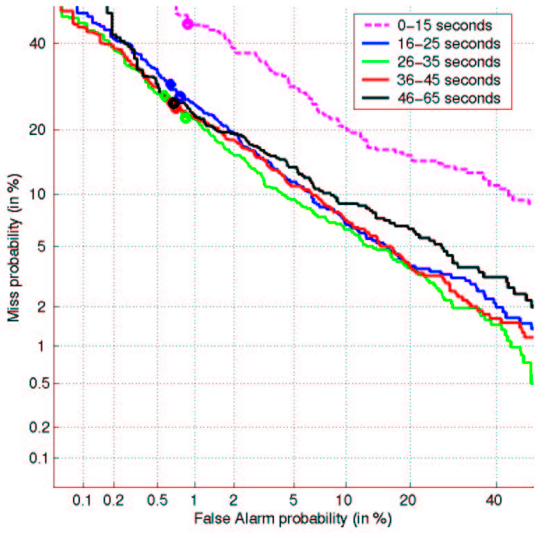

Figure 5: DET curves by duration for one well-performing system in the 2002 one-speaker detection task.

5.3 Pitch

Intuitively the stability of a speaker's voice aids performance in speaker recognition. Many factors have the ability to destabilize a speaker's voice. Looking specifically at the average pitch of speakers' voices, depending on the time of day, or whether or not they have a cold, the average pitch is easily affected.

In the NIST evaluations the speaker's model is created from approximately two minutes of speech. The average pitch of this training model can be estimated [9], and compared with the similarly estimated pitch of each segment speaker, including in particular the segments containing the model speaker that were recorded at different times in different conditions. We find that the speaker's pitch can change drastically. To analyze how this difference in pitch affects performance, we look at the set of target trials (true speaker tests) where the pitch of the test segment speaker is close to the pitch of the model speaker and compare this performance to that where the pitch values are far apart.

Figure 6 shows the results for one system when the target trials are limited by such pitch closeness conditions. NIST defined two categories, the 25% of the trials that were closest in average pitch between the model and segment, and the 25% of the trials that were furthest in average pitch, using

measurements on a log scale. The system was one included in the 1999 evaluation.

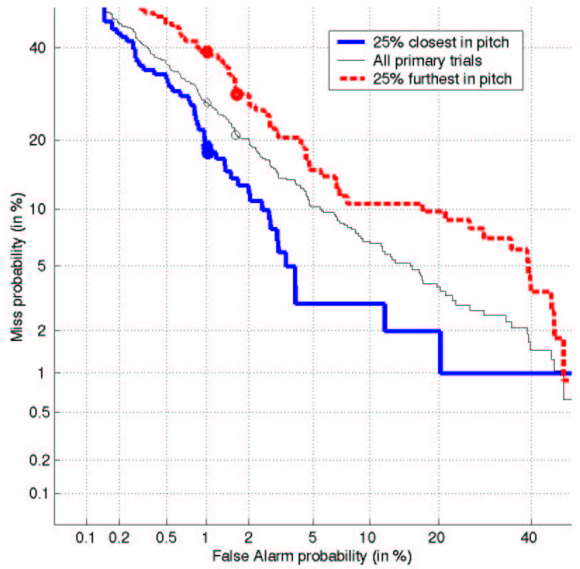

Figure 6: DET plot showing the effect of target trial pitch closeness on performance for one system in the 1999 evaluation.

5.4 Landline versus Cellular Transmissions

The past few years have brought an increasing interest in the processing of cellular data for both speech and speaker recognition. The two most recent Switchboard collections have been collections of primarily cellular data. The first collection was used as a secondary test set in 2001, while the second collection was used as the main evaluation data in 2002.

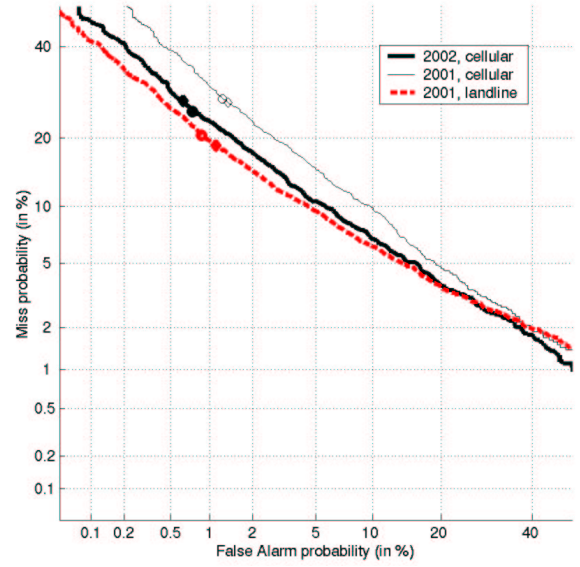

Figure 7: DET plot showing the difficulty of cellular vs. landline test data, as well as the progress made in one year of working with cellular data.

Figure 7 presents DET plots of the best performing systems in the 2001 landline, 2001 cellular, and the 2002 cellular evaluation sets.

If the two cellular test sets are of comparable difficulty, and other comparisons suggest that they are, then figure 7 shows some real improvement in the best system performance between 2001 and 2002. The 2001 curves show that the cellular test sets are measurably more difficult than the landline data. This comparison, however, rather understates the difference in relative difficulty of landline and cellular data. This is because the landline data was selected so that the target trials always involved different handsets in training and test data, but the collection protocol for cellular didn't permit this, and in most target trials the training and test handsets are the same.

5.5 Other conditions previously reported

Other conditions that have been reported elsewhere regarding results of assessing speaker recognition technology through the NIST series of evaluations are included in this section.

5.5.1 Training conditions

In the 1997 evaluation, there were three training conditions:

- One-session: Training data comes from one conversation, and therefore one telephone handset.
- One-handset: Training data comes from two conversations, but from a single telephone number (and presumably one telephone handset)
- Two-handset: Training data comes from two conversations of differing phone numbers (and presumably two telephone handsets)

At the 1998 RLA2C workshop, NIST reported that, to a measurable degree, the more varied the training data, the better the performance in speaker recognition [10]. At that time, when restricting our analysis to the different number tests for each training condition, at a 5% miss rate the best system had a 10% false alarm rate with two-handset training, a 20% false alarm rate with one-handset training, and a 35% false alarm rate with one-session training.

As future speaker recognition evaluations were becoming more complex, with an array of tasks and training conditions, NIST used these findings to decide to concentrate on increasingly difficult problems. In 1999 only the two-handset training condition was offered. One-session training has been the focus since the 2000 evaluation.

5.5.2 Gender

When separating performance by male and female tests, NIST has found that the better performing systems generally had somewhat better performance with male data than with female data, as was the case in 1997, 1999 and, to a lesser degree, in 1998 and 2000. Moving to cellular data in 2001 and 2002, there was virtually no difference in performance by gender.

5.5.3 Same number versus different number tests

It has been shown that speaker recognition performance improves if the training of each model speaker comes from the same telephone handset as the test segments of the same speaker, referred to here as same number tests.

In the 1999 evaluations using landline data, NIST observed that at a 5% miss rate, performance for same number tests had a 1.5% false alarm rate, while that for different number tests had a 10% false alarm rate.

Recent evaluations using cellular data have included primarily same number tests. Most speakers used their own single cell phone. There were thus not many, different number, all cellular, target trials, though trials involving a landline phone were often different number. Therefore, conditioning on same number versus different number tests was not attempted.

5.5.4 Handset Types, matched and mismatched

Most standard landline telephone handset microphones are of either the carbon-button or electret type. We observed in early evaluations that the handset types used, both in the training and test segments, can greatly influence recognition.

MIT-Lincoln Laboratory, a participant in every NIST Speaker Recognition evaluation, developed an automatic handset labeller [11], which uses the telephone speech signal from one channel to assign a likelihood that the signal is from a carbon-button handset. This likelihood is converted into a hard decision (carbon or electret). Although less than perfect, the accuracy of the handset labeller is believed to be very high. The hard decisions for all landline training and test data have been given as side information since 1997.

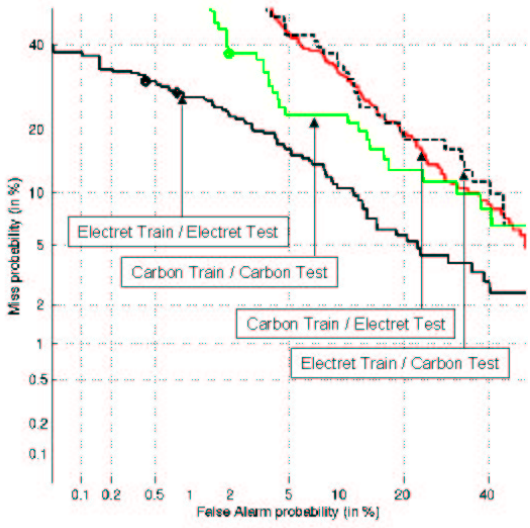

Figure 8: Performance as a function of training/test handset type. Performance for one system on different number tests for each combination of training and test handset types. Performance improves when the types match and is superior for electret handsets.

Figure 8 shows the variation in performance for different combinations of training and test-segment handset types for a well performing system in the 2001 evaluation on landline data.

6. EXTENDED DATA TASK

The results of the extended data task in 2001 and 2002 were quite dramatic. In 2002 four sites worked either independently or in teams on this task. Several techniques were explored, and among the four sites, 16 systems were submitted. (When a site submits more than one system for a task, it is required to designate which system is considered its primary one).

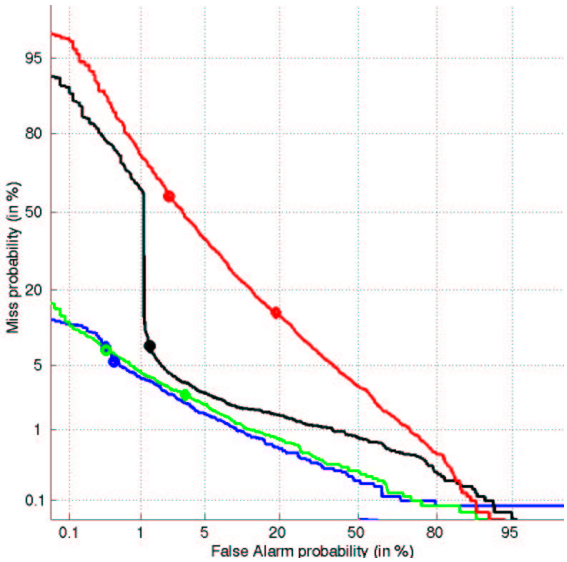

Figure 9: DET plot of the four primary systems for the 2002 extended data task. Conventional one-speaker detection systems had an equal error rate of just under 10%.

Figure 9 shows a DET plot for each site's primary system when the trials were restricted to the primary condition, which were:

- Trials with segment speakers (for both target and non-target trials) restricted to those with at least 17 conversation sides (allowing 17 or more test segments for these speakers).
- Target and non-target trials involving models trained with 8 conversation sides.

There were over 59,000 trials in the 2002 extended data task. These restriction provided 6,127 target trials involving 291 speakers. There were also over 6,900 non-target trials.

At a 5% miss rate, the system with the best performance has a false alarm rate of less than 1%!

7. MEASURING PROGRESS

The primary goal of the NIST speaker recognition evaluations is to measure and track the progress of the technology.

Progress is often difficult to show, because of the test set variability and change in focus and data sets.

Figure 10 shows the history of speaker recognition performance in the NIST speaker recognition evaluations. Shown are the top systems from each year for the basic task of one-speaker detection on landline data, the foundation on which the technology has been built.

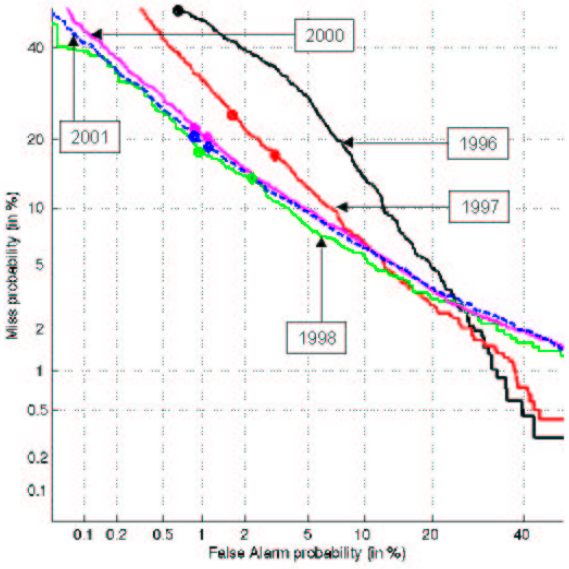

Figure 10: History of the NIST one-speaker detection results on landline data when limiting test trials to similar conditions (1-session training, 30 second average test segment duration, electret handsets). Steady improvement is shown from 1996 through 1998. 1999 is not shown because only two-handset training was offered. The 2000 and 2001 test sets perhaps represent a slightly more difficult task due to the use of varied length tests segments (15 to 45 second durations).

Note that for this figure, instead of plotting each year's primary condition results, we have preferred to plot "similar conditions" from the corresponding evaluations. For example, in years when "same number tests" were of primary interest, there was a secondary condition of "different number tests". Because of the large performance gap between same number and different number tests, "different number tests" results are used throughout to track progress.

8. CONCLUSIONS

The extended data task has led to some dramatic progress in speaker detection when large amounts of both training and test data are available. Work in the NIST evaluations and at a Johns Hopkins workshop in the summer of 2002 [12] has shown that automatic systems can achieve very high performance levels by combining various levels of information from the speech signal and taking advantage of progress in automatic speech recognition. This is analogous to the capabilities human have to identify familiar speakers. Future NIST evaluations may confirm and enhance these promising results.

Progress on the basic one-speaker detection problem with lesser amounts of training and test data has been more limited, however. The greater prevalence of cellular handsets, moreover, has added to the difficulty of the problem and the factors that affect cellular performance, including handset types and encoding systems, require further study. Future NIST evaluations will attempt to address these issues.

Current plans call for the annual NIST Speaker Recognition evaluations [13] to continue. The Linguistic Data Consortium has plans to collect further conversational data, both landline and cellular, to support future research.

NIST would be interested in other appropriate conversational data sources, including voices recorded over the Internet, particularly in languages other than English. The evaluations remain open to all sites interested in participating in accordance with the evaluation rules.

9. REFERENCES

1. P. Price, W. Fisher, J. Bernstein and D. Pallett, http://www.ldc.upenn.edu/Catalog/LDC93S3A.html, "Resource Management Complete Set 2.0"
2. NIST Rich Transcription Evaluation, 2003, http://www.nist.gov/speech/tests/rt/rt2003/
3. Doddington, G., "Speaker Recognition based on Idiolectal Differences between Speakers", Eurospeech 2001, Vol. 4, pp. 2521-2524
4. Linguistic Data Consortium, http://www.ldc.upenn.edu//Catalog/SID.html
5. An instantiation of the BBN Byblos speech-to-text system was installed and operated by NIST, to generate the automatically generated transcripts. It was not an evaluation system, optimized for ASR accuracy.
6. "DETware scoring software", http://www.nist.gov/speech/tools/DETware_v2.1.tar.gz
7. Swets, John A, ed., "Signal Detection and Recognition by Human Observers", John Wiley & Sons, Inc., pp. 611-648, 1964.
8. Martin, Alvin, et al., "The DET Curve in Assessment of Detection Task Performance", 1997 Eurospeech proceedings, Vol. 4, pp. 1899-1903
9. HTK Entropic's software, "get_F0", http://htk.eng.cam.ac.uk/
10. Przybocki, M., Martin, A., "NIST Speaker Recognition Evaluation – 1997", 1998 RLA2C
11. Quatieri, T., Reynolds, D., O'Leary, G., "Magnitude-Only Estimation of Hnadset Nonlinearity with Application to Speaker Recognition", Proceedings ICASSP (1998), pp. 745-748

12. "SuperSID: Exploiting High-Level Information for High-Performance Speaker Recognition", The Center for Language and Speech Processing, 2002 Summer Workshop, http://www.clsp.jhu.edu/ws2002/groups/superisd/

13. NIST Speaker Recognition evaluations – General Information, at http://www.nist.gov/speech/tests/spk/index.htm